Estimating 2D Camera Motion with Hybrid Motion Basis (Supplementary Materials)

Anonymous ICCV submission

Paper ID 3239

1. Summarize

This supplementary material provides extended details and evaluations for the paper *Estimating 2D Camera Motion with Hybrid Motion Basis*. The key contributions and additional information are summarized as follows:

- The visual qualitative in HTML format and videos (README.html).
- Technical details and comprehensive related work.
- Experiments on extra benchmarks.
- Technique details for the Application.

2. Related works

2.1. Rotation-Only Homography

Rotation-only homography usually computes from the gyroscope, which records the 3D camera movements, is widely used in various mobile imaging devices, providing accurate measurements of device rotation. Many applications, such as video stabilization [22], optical image stabilization (OIS) [25], image deblurring [38], simultaneous localization and mapping (SLAM) [20], ego-motion estimation [4], gesture-based user authentication on mobile devices [17], image alignment with OIS calibration [32], and human gait recognition [51], extensively utilize gyroscope data. Due to its high accuracy, gyroscope information serves as a reliable source for various tasks.

2.2. Homography Methods

Traditional homography estimation typically follows three stages: feature detection using algorithms such as SIFT [36] or ORB [41], correspondence matching [8], and outlier rejection techniques like RANSAC [13]. Recent advances in learning-based feature detection and matching, including LIFT [47], SuperPoint [10], and SOSNet [43], have improved the robustness of homography estimation. Additionally, enhanced outlier rejection methods such as MAGSAC [2] and MAGSAC++ [3] have increased stability in challenging scenarios involving multiple planes, parallax, and dynamic foregrounds. Optimization-based approaches [7, 11], such as those derived from Lucas-Kanade or sum of squared differences, iteratively refine homography parameters from an initial estimate. Deep learning methods have further advanced homography estimation, beginning with supervised approaches [9] that rely on synthetic image pairs. More recent methods can be categorized into supervised [6, 26, 42] and unsupervised [19, 23, 39, 46] frameworks. Unsupervised techniques have gained popularity due to their label-free training strategies. For example, CAHomo [50] utilizes a self-guided mask to highlight key feature points, while BasesHomo [46] constrains the rank of feature maps by learning an 8-dimensional motion basis for improved estimation. HomoGAN [19] introduces a Generative Adversarial Network (GAN) loss to identify the dominant plane and integrates a Transformer encoder for coarse-to-fine refinement. Additionally, SCPNet [52], McNet [53], and InterNet [48] explore cross-modal homography estimation. Despite these advancements, homography remains a single-plane parametric model, limiting its ability to fully represent complex camera motion. To overcome this constraint, we introduce a hybrid motion-basis representation designed to model multi-plane, non-linear motion more effectively.

2.3. Multi Homography Methods

Mesh-based image warping is commonly applied to scenes with multiple planes, where each mesh cell provides a local homography. Gao *et al.* [14] introduced a dual homography model that estimates separate homographies for the distant plane and the ground plane. The As-Projective-As-Possible (APAP) approach [49] computes mesh warps that follow a global projective transformation but allow local deviations for each mesh grid. The Bundled Paths approach [30] solves an as-similar-as-possible warp [21] based on matched features, then estimates local homographies for each mesh grid from the deformed mesh. MeshFlow [31] uses motion vectors from matched feature correspondences to guide the mesh warp. In deep learning-based methods, MeshCAHomo [33] estimates multiple meshes at different resolutions and merges them for the final warp, while Mesh-

BasesHomo [34] learns mesh flow by combining motion bases within each grid. MeshHomoGAN [35] incorporates a planarity-aware mechanism for improved local homography estimation. However, these motion representations still rely on homography, limiting their ability to represent complex camera motion. We propose a solution that designs multiple non-linear motion bases and combines them using a neural network.

2.4. Image Matching Dataset

Image matching datasets encompass both optical flow and homography datasets. Several well-known datasets are available for optical flow estimation and benchmarking, including FlyingChairs [12], MPI-Sintel [5], KITTI 2012 [15], KITTI 2015 [37], and GOF [27]. Additionally, some modern approaches can generate precise optical flow ground-truth labels. For example, Infinigen [40] is a procedural tool that synthesizes highly realistic 3D natural environments using Blender, while Kubric [16] also utilizes Blender to produce large-scale datasets featuring annotated photo-realistic scenes. Nevertheless, this study primarily focuses on homography datasets.

Among the most widely used datasets for homography estimation are MegaDepth, HPatches, CAHomo, GF4, and GHOF, which serve as benchmarks for both training and evaluating deep learning models. MegaDepth [29] is built using multi-view internet photo collections and is processed with advanced Structure-from-Motion (SfM) and Multi-View Stereo (MVS) techniques. HPatches [1] consists of image sequences captured under different lighting and viewpoint variations. CAHomo [50] is designed for assessing homography estimation in challenging conditions, making it the first dataset to introduce a diverse range of such scenarios. The GF4 dataset [32] uniquely integrates gyroscope data with video frames, providing sparse annotations. Furthermore, GHOF [28] merges gyroscope data with video sequences, enabling the evaluation of both homography and optical flow estimation methods.

Despite the availability of these datasets, there is a notable absence of high-quality datasets for camera motion estimation. To address this gap, we propose a novel benchmark via masking the existing optical flow testset.

3. Experiment

3.1. CAHomo Dataset

The CA-Unsupervised dataset [50] consists of 800,000 training image pairs and 4,200 test pairs, captured from five distinct real-world environments. These range from common scenes to challenging conditions such as low illumination, minimal texture, and varying foreground scales, where homography estimation becomes particularly complex. The training set is unlabeled, containing only consecutive frame pairs, while the test set includes 6–10 manually annotated keypoints. The first six keypoints are placed on background regions or dominant planes, whereas the last four are optionally assigned to foreground planes, facilitating precise performance evaluation.

3.2. GHOF Dataset

The Gyroscope-Homography-Optical-Flow (GHOF) dataset [28] is designed for both homography and optical flow analysis, incorporating gyroscope data from non-optically stabilized (non-OIS) cameras. It comprises 10,000 training samples and 256 test pairs, covering five scene categories: regular (RE), foggy (FOG), low-light (LL), rainy (RAIN), and snowy (SNOW). Unlike CAHomo, GHOF presents additional challenges due to significant parallax and extreme foreground-background variations. The test set provides 5–8 sparse, annotated correspondences on background regions for evaluation.

Existing datasets, such as CAHomo and GHOF, primarily focus on background motion or dominant plane alignment, representing only a subset of full camera motion. While CAHomo can potentially evaluate foreground motion, it has two main limitations: (1) its sparse annotations fail to capture camera motion and its smoothness adequately, and (2) the annotation process described in [50] is unsupervised, leading to potential inaccuracies in keypoint matching. To address these shortcomings, we leverage the GHOF dataset while masking dynamic objects to isolate camera-specific motion within static scenes, introducing the novel GHOF-Cam Benchmark.

3.3. Implementation Details

Our proposed CamFlow builds upon HomoGAN [19], a state-of-the-art unsupervised framework for homography estimation. We modify its token blocks to predict N weights and replace the original eight motion bases with our N hybrid bases. Optimization is performed using the Adam optimizer [24] with a learning rate of 1.0×10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. The model is trained with a batch size of 16, using input images of 600×800 pixels, cropped into 384×512 patches, over five epochs on 800,000 image pairs from CAHomo [50]. The learning rate decays by a factor of 0.8 after each epoch. Training is completed in 14 hours using four NVIDIA A800 GPUs or in two days using eight NVIDIA 2080Ti GPUs. The weight of the negative log-Laplace motion loss is set to 0.1.

3.4. Other Multi-Plane Benchmarks

Beyond evaluating background motion (homography) and our proposed GHOF-Cam, we also conduct experiments on existing meshflow-based benchmarks [45], which, despite their limited accuracy, offer useful comparisons. To this end, we experiment with foreground sparse annotations and

Table 1. Point matching errors of our MeshHomoGAN compared to existing methods for mesh-based homography estimation.

		RE	LT	LL	SF	LF	Avg
1)	$\mathcal{I}_{3\times3}$	7.81	7.87	7.49	8.34	4.14	7.13
2)	APAP	1.59	2.72	1.75	1.70	2.10	1.97
3)	ANAP	1.67	3.14	1.91	2.38	2.72	2.36
4)	MeshFlow	0.46	1.04	1.06	1.09	1.36	1.00
5)	Unsupervised	0.50	1.3	1.09	1.26	1.36	1.11
6)	MeshCAHomo	0.53	1.23	1.03	1.17	0.96	0.98
7)	MeshBasesHomo	0.32	0.91	0.67	0.48	0.74	0.62
8)	MeshHomoGAN	0.22	0.56	0.43	0.39	0.73	0.47
9)	Ours	0.26	0.58	0.57	0.56	0.86	0.56

report the results in Table 1.

4. Application

We recorded several real-world videos and applied offline digital video stabilization (DVS) methods based on Liu et al. [30]. The resulting videos, included in the supplementary materials, demonstrate that our approach significantly enhances visual performance.

Traditional DVS comprises three main steps: (1) physical motion estimation, (2) motion optimization, and (3) image rendering. In our method, we retained the latter two stages while replacing the first with deep homography estimation using CamFlow. Our hybrid motion modeling enables more accurate camera motion capture. Additionally, with flexible loss functions, we enable test-time adaptation on unseen videos by maintaining photometric loss while replacing pseudo-motion labels with IMU-based gyroscope motion.

5. Test-Time Adaptation

5.1. IMU Pseudo Labels

Gyroscope sensors capture the relative 3D rotation of a camera over time. By converting gyroscope readings into a 2D motion field, frames can be aligned [27]. This is also beneficial for correcting rolling shutter (RS) effects [44]. Given the 3-axis angular velocities (roll $\mathbf{v_r}$, pitch $\mathbf{v_p}$, and yaw $\mathbf{v_y}$) and the time interval Δt, the angular rotations are computed as:

$$\angle\mathbf{r} = \mathbf{v_r} \cdot \Delta t, \quad \angle\mathbf{p} = \mathbf{v_p} \cdot \Delta t, \quad \angle\mathbf{y} = \mathbf{v_y} \cdot \Delta t. \tag{1}$$

Using the Rodrigues formula, we convert these into a rotation matrix $\mathbf{R}(\Delta t) \in SO(3)$, which is further transformed into a homography matrix $\mathbf{H}$ [18].

5.2. Fine-Tuning on Unseen Videos

For novel videos, we extract the Gyro Field and Intra Gyro Field between consecutive frames, applying IGF for RS correction. During fine-tuning, we replace pseudo-motion labels in our hybrid loss function:

$$\ell_{overall} = \ell_{NLL_p} + \mathbf{w} \times \frac{\left|\ell_{NLL_p}\right|}{\left|\ell_{NLL_m}\right|} \cdot \ell_{NLL_m}. \tag{2}$$

This strategy effectively leverages the stability of sensor data while mitigating its rotational limitations, leading to enhanced video stabilization and improved visual quality.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5173–5182, 2017. 2

[2] Daniel Barath, Jiri Matas, and Jana Noskova. Magsac: marginalizing sample consensus. In *Proc. CVPR*, pages 10197–10205, 2019. 1

[3] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1304–1312, 2020. 1

[4] Michael Bloesch, Sammy Omari, Péter Fankhauser, Hannes Sommer, Christian Gehring, Jemin Hwangbo, Mark A Hoepflinger, Marco Hutter, and Roland Siegwart. Fusion of optical flow and inertial measurements for robust egomotion estimation. In *Proc. IROS*, pages 3102–3107, 2014. 1

[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proc. ECCV*, pages 611–625, 2012. 2

[6] Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *Proc. CVPR*, pages 1879–1888, 2022. 1

[7] Che-Han Chang, Chun-Nan Chou, and Edward Y Chang. Clkn: Cascaded lucas-kanade networks for image alignment. In *Proc. CVPR*, pages 2213–2221, 2017. 1

[8] Padraig Cunningham and Sarah Jane Delany. K-nearest neighbour classifiers-a tutorial. *ACM Computing Surveys (CSUR)*, 54(6):1–25, 2021. 1

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 1

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. CVPRW*, pages 224–236, 2018. 1

[11] Tianjiao Ding, Yunchen Yang, Zhihui Zhu, Daniel P Robinson, René Vidal, Laurent Kneip, and Manolis C Tsakiris. Robust homography estimation via dual principal component pursuit. In *Proc. CVPR*, pages 6080–6089, 2020. 1

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proc. ICCV*, 2015. 2

[13] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[14] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *Proc. CVPR*, pages 49–56, 2011. 1

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. CVPR*, pages 3354–3361, 2012. 2

[16] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3761, 2022. 2

[17] Dennis Guse and Benjamin Müller. Gesture-based user authentication on mobile devices using accelerometer and gyroscope. In *Informatiktage*, pages 243–246, 2012. 1

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[19] Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, Qijun Zhao, and Shuaicheng Liu. Unsupervised homography estimation with coplanarity-aware gan. In *Proc. CVPR*, pages 17663–17672, 2022. 1, 2

[20] Weibo Huang and Hong Liu. Online initialization and automatic camera-imu extrinsic calibration for monocular visual-inertial slam. In *Proc. ICRA*, pages 5182–5189, 2018. 1

[21] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM Trans. Graphics*, 24(3):1134–1141, 2005. 1

[22] Alexandre Karpenko, David Jacobs, Jongmin Baek, and Marc Levoy. Digital video stabilization and rolling shutter correction using gyroscopes. *CSTR*, 1(2):13, 2011. 1

[23] Dewi Endah Kharismawati, Hadi Ali Akbarpour, Rumana Aktar, Filiz Bunyak, Kannappan Palaniappan, and Toni Kazic. Cornet: Unsupervised deep homography estimation for agricultural aerial imagery. In *Proc. ECCV*, pages 400–417, 2020. 1

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[25] Fabrizio La Rosa, Maria Celvisia Virzì, Filippo Bonaccorso, and Marco Branciforte. Optical image stabilization (ois). 2015. STMicroelectronics. Available online; accessed on 31 October 2015. 1

[26] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In *Proc. CVPR*, pages 7652–7661, 2020. 1

[27] Haipeng Li, Kunming Luo, and Shuaicheng Liu. Gyroflow: Gyroscope-guided unsupervised optical flow learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12869–12878, 2021. 2, 3

[28] Haipeng Li, Kunming Luo, Bing Zeng, and Shuaicheng Liu. Gyroflow+: Gyroscope-guided unsupervised deep homography and optical flow learning. *arXiv preprint arXiv:2301.10018*, 2023. 2

[29] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2

[30] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. Bundled camera paths for video stabilization. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013. 1, 3

[31] Shuaicheng Liu, Ping Tan, Lu Yuan, Jian Sun, and Bing Zeng. Meshflow: Minimum latency online video stabilization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 800–815. Springer, 2016. 1

[32] Shuaicheng Liu, Haipeng Li, Zhengning Wang, Jue Wang, Shuyuan Zhu, and Bing Zeng. Deepois: Gyroscope-guided deep optical image stabilizer compensation. *IEEE Trans. on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2021.3103281, 2021. 1, 2

[33] Shuaicheng Liu, Nianjin Ye, Chuan Wang, Jirong Zhang, Lanpeng Jia, Kunming Luo, Jue Wang, and Jian Sun. Content-aware unsupervised deep homography estimation and its extensions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(3):2849–2863, 2022. 1

[34] Shuaicheng Liu, Yuhang Lu, Hai Jiang, Nianjin Ye, Chuan Wang, and Bing Zeng. Unsupervised global and local homography estimation with motion basis learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(6): 7885–7899, 2023. 2

[35] Shuaicheng Liu, Mingbo Hong, Yuhang Lu, Nianjin Ye, Chunyu Lin, and Bing Zeng. Unsupervised global and local homography estimation with coplanarity-aware gan. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[36] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[37] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proc. CVPR*, pages 3061–3070, 2015. 2

[38] Janne Mustaniemi, Juho Kannala, Simo Särkkä, Jiri Matas, and Janne Heikkila. Gyroscope-aided motion deblurring with deep networks. In *Proc. WACV*, pages 1914–1922, 2019. 1

[39] Ty Nguyen, Steven W Chen, Shreyas S Shivakumar, Camillo Jose Taylor, and Vijay Kumar. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robotics and Automation Letters*, 3(3):2346–2353, 2018. 1

[40] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, et al. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 2

[41] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proc. ICCV*, pages 2564–2571, 2011. 1

[42] Ruizhi Shao, Gaochang Wu, Yuemei Zhou, Ying Fu, Lu Fang, and Yebin Liu. Localtrans: A multiscale local transformer network for cross-resolution homography estimation. In *Proc. ICCV*, pages 14890–14899, 2021. 1

[43] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proc. CVPR*, pages 11016–11025, 2019. 1

[44] Zhanglei Yang, Haipeng Li, Mingbo Hong, Bing Zeng, and Shuaicheng Liu. Single image rolling shutter removal with diffusion models. *arXiv preprint arXiv:2407.02906*, 2024. 3

[45] Nianjin Ye, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Jue Wang, and Yongqing Cui. Deepmeshflow: Content adaptive mesh deformation for robust image registration. *arXiv preprint arXiv:1912.05131*, 2019. 2

[46] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *Proc. ICCV*, pages 13117–13125, 2021. 1

[47] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Proc. ECCV*, pages 467–483, 2016. 1

[48] Junchen Yu, Si-Yuan Cao, Runmin Zhang, Chenghao Zhang, Jianxin Hu, Zhu Yu, Beinan Yu, and Hui-liang Shen. Internet: Unsupervised cross-modal homography estimation based on interleaved modality transfer and self-supervised homography prediction. *arXiv preprint arXiv:2409.17993*, 2024. 1

[49] Julio Zaragoza, Tat-Jun Chin, Michael S Brown, and David Suter. As-projective-as-possible image stitching with moving dlt. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2339–2346, 2013. 1

[50] Jirong Zhang, Chuan Wang, Shuaicheng Liu, Lanpeng Jia, Nianjin Ye, Jue Wang, Ji Zhou, and Jian Sun. Content-aware unsupervised deep homography estimation. In *Proc. ECCV*, pages 653–669, 2020. 1, 2

[51] Rong Zhang, Christian Vogler, and Dimitris Metaxas. Human gait recognition. In *Proc. CVPRW*, pages 18–18, 2004. 1

[52] Runmin Zhang, Jun Ma, Si-Yuan Cao, Lun Luo, Beinan Yu, Shu-Jie Chen, Junwei Li, and Hui-Liang Shen. Scpnet: Unsupervised cross-modal homography estimation via intra-modal self-supervised learning. In *European Conference on Computer Vision*, pages 460–477. Springer, 2024. 1

[53] Haokai Zhu, Si-Yuan Cao, Jianxin Hu, Sitong Zuo, Beinan Yu, Jiacheng Ying, Junwei Li, and Hui-Liang Shen. Mcnet: Rethinking the core ingredients for accurate and efficient homography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25932–25941, 2024. 1